A STEP AHEAD FOR GENERATION OF EXPLAINABLE EMBEDDINGS USING FEATURE MAXIMIZATION PRINCIPLE

Jean-Charles LAMIREL

Synalp-Team-LORIA

University of Strasbourg

SAMM – Paris 1

WISELAB – DUT DALIAN (Sea-sky Scholar)

OCTA 2023

Introduction

Data mining is a machine learning domain which raises difficult challenges mainly because of high dimensional data and large datasets:

- Facing with learning or mining model evaluation
- **Facing with distance ambiguities or inefficiency**
- Facing with multiple data representations
- **Facing with synthetizing and visualizing mining results**
- Facing with potentially evolving data

Does it exist a federating approach that can globally deal with such problem ?

Could the approach be exploited for visualization and management of big graphs (including embeddings) ?

Presentation plan

- Principle of feature maximization metric
- Exploiting F-max for complex data analysis
  - Extension of feature maximization metric for community role detection
  - Sparse graph embedding
  - Sparse word embedding based on community roles
- Perspectives

A new metric for high dimensional data management : the feature maximization metric (F-max)

An alternative to usual metrics : the feature maximization metric [Lamirel 08]

Let us consider a set of clusters C resulting from a clustering method applied on a set of data D represented with a set of descriptive features F, feature maximization is a metric which favors clusters with maximum *Feature F-measure* which represents the harmonic mean between :

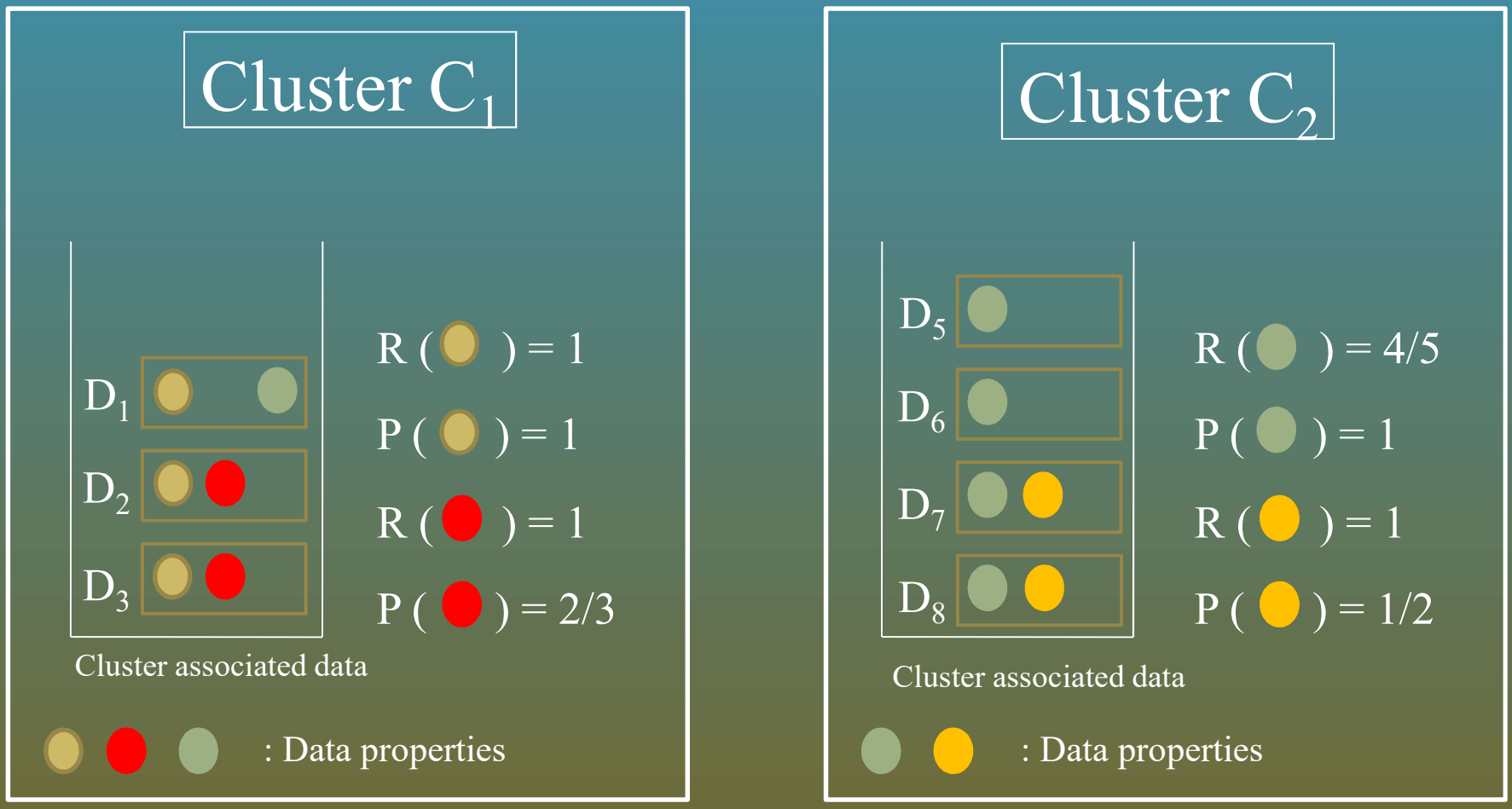
Feature Recall $FR_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{c' \in C} \sum_{d \in c'} W_d^f}$ $\equiv P(c|f)$

Feature Predominance $FP_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{f' \in F^c, d \in c} W_d^{f'}}$ $\equiv P(f|c)$

A **maximized cluster feature** is a feature whose *Feature F-measure* is **maximized** by the cluster members (i.e. data).

Measures based on data description space

A simplified view with binary Recall and Precision [Lamirel 04]



The R, P, F–measure criteria are independent of the clustering method (symbolic equivalence).

Feature maximization metric

Extended use

In machine learning feature maximization metric proved to have very various use, like:

- Extraction of association rules [Al Shehabi 2006]
- Optimizing learning [Attik 2006]
- Cluster labeling and cluster content mining [Lamirel 2008]
- Detecting incoherent clustering results [Lamirel 2010a]
- Substituting to distance in clustering [Lamirel 2011] [Falk 2012]
- Efficient feature selection for supervised classification (dealing with class imbalance, noise, …) [Lamirel 2016a]
- Evaluating/Correcting initial clustering results [Lamirel 2016b]
- Substituting to distance between distribution [Olteanu 2019]
- SNA analysis and data synthesis and summarization [Lamirel 2020]

A simple example

- We consider a sample of **Men (M)** and **Women (F)** for which we measure **Hair_length** and **Shoes_size** and **Nose_size**

Shoes _size	Hair _length	Nose _size	Class
9	5	5	M
9	10	5	M
9	20	6	M
5	15	5	F
6	25	6	F
5	25	5	F

A simple example

- We compute the Feature Recall (FR) and the Feature Precision (FP) and the Feature F-measure (FF) for each class and each feature and each class

Shoes _size	Hair _length	Nose _size	Class
9	5	5	M
9	10	5	M
9	20	6	M
5	15	5	F
6	25	6	F
5	25	5	F

$$FR(S,M) = 27/43 = 0.62$$

$$FP(S,M) = 27/78 = 0.35$$

$$FF(S,M) = \frac{2(FR(S,M) \times FP(S,M))}{FR(S,M) + FP(S,M)} = 0.48$$

A simple example

- We compute the average marginal values of Feature F-measure by feature (local) and the overall Feature F-measure for each class and each feature and each class

	F(x,M)	F(x,F)	$\overline{F(x,.)}$
Hair_ length	0.39	0.66	0.53
Shoes_ size	0.48	0.22	0.35
Nose_ size	0,3	0,24	0,27

$\overline{F(.,.)}$
0.38

The features whose Feature F-measure is under the global Feature F-mesure average are removed

⇒ Nose_size is removed

The remaining (i.e. selected) features whose F-measure is over marginal average in one class are considered as active in this class

=> Shoes_size is active in Men class

=> Hair_length is active in Women class

A simple example

- The contrast factor highlights the degree of activity/passivity of selected features relatively to their marginal Feature F-measure average in the different classes

	F(x,M)	F(x,F)	$\overline{F(x,.)}$
Hair _length	0.39	0.66	0.53
Shoes _size	0.48	0.22	0.35

	C(x,M)	C(x,F)
Hair _length	0.39/0.53	0.66/0.53
Shoes _size	0.48/0.35	0.22/0.35

	C(x,M)	C(x,F)
Hair _length	0.74	1.25
Shoes _size	1.37	0.63

The constrast can be seen as a function that will tend to:

1. **Overlength the Hairs of Women**
2. **Oversize the Shoes of Men**
3. **Underlength the Hairs of Men**
4. Undersize the Shoes of Women

A simple example

- The contrast is applied on the data in order to modify the feature weights depending on the data class

Shoes _size	Hair _length	Class
9	5	M
9	10	M
9	20	M
5	15	F
6	25	F
5	25	F

Original data

Shoes _size	Hair _length	Class
12,33	3.7	M
12,33	7.4	M
12,33	14.8	M
3.15	18.75	F
3,78	31.25	F
3.15	31.25	F

Contrasted data

Data contrast can change the organization of the data in the description space in a non linear way

A simple example

- The magnification factor (k) can enhance the contrast to facilitate classification in complex cases

Shoes _size	Hair _length	Class
12,33	3.7	M
12,33	7.4	M
12,33	14.8	M
3.15	18.75	F
3,78	31.25	F
3.15	31.25	F

Contrasted data (k = 1)

Shoes _size	Hair _length	Class
28.30	1.59	M
28.30	3.19	M
28.30	6.37	M
0.99	36.47	F
1.20	60.79	F
0.99	60.79	F

Contrasted data (k = 4)

Magnification is a non linear transformation (enhance non-linearity)

F-max feature selection and contrasting (FMC)

Explanation capabilities – ex: 20 Newgroups collection

Selected features with highest contrast can be used to provide explanation on class content

rec.autos	rec.motorcycles	sci.electronics	comp.graphics	sci.space
14.40 car	14.89 ride	13.69 circuit	12.22 imag	14.51 launch
14.01 ford	14.09 dog	13.42 wire	11.67 viewer	14.24 moon
10.54 auto	13.40 dod	11.89 outlet	11.03 graphic	13.53 mission
9.98 alarm	11.62 helmet	10.35 ham	9.88 render	12.87 space
9.79 shift	9.39 lean	9.92 concret	9.69 manipul	12.41 solar
9.74 mileag	9.23 chain	9.13 relai	8.50 pub	12.29 planet
9.20 oil	9.22 cage	9.03 neutral	8.34 plot	11.81 satellit
8.64 gear	9.11 cit	8.82 tone	8.34 gif	11.79 atmospher
7.96 tire	8.89 drink	8.81 ground	7.77 crop	11.17 sky
7.87 transmiss	8.49 newbi	8.61 led	7.74 format	10.82 henri

sci.med	talk.politics.guns	talk.politics. mideast	soc.religion. christian	alt.atheism
14.67 patient	14.43 gun	12.07 occupi	10.66 bless	8.10 keith
13.24 medic	10.18 cdt	11.19 villag	10.54 rutger	7.84 societ
12.60 doctor	10.08 accident	11.02 soldier	9.58 sin	7.78 moral
12.00 food	9.51 revolv	10.75 territori	9.05 marri	7.34 belief
11.65 medicin	9.18 compound	9.62 israel	8.93 church	7.30 vice
11.56 treatment	8.57 semi	9.14 border	8.76 spirit	7.20 instinct
11.45 clinic	8.38 fire	9.00 shout	8.35 mari	6.93 evolut
11.41 infect	8.19 raid	8.50 turkei	8.22 easter	6.74 jon
11.30 cure	8.13 assault	8.45 arab	8.07 pope	6.74 bake
10.93 Gordon	8.07 weapon	8.39 greec	8.06 geneva	6.72 speci

Classification with FMC

Deft challenge [Lamirel - JADT 2014]

- Dataset of extracts of talk of CHIRAC et MITTERAND presidents:
  - 73255 sentences of Chirac,
  - 12320 sentences of Mitterrand.
- Best results till now on that dataset : 88% accuracy (almost 16850 bilateral errors) by LIA.
- Result with feature maximization : 99,999% accuracy (12 unilateral errors)
  - Strong feature selection (~50000 → ~5000)
  - Extra-light NLP preprocessing,
  - No lemmatization is needed,
  - Stop words are kept and proof to be useful for analysis.

CHIRAC	MITTERAND
1.930810 partenariat	1.881835 douze
1.858265 dynamisme	1.852007 est-ce
1.811123 exigence	1.800091 eh
1.775048 compatriotes	1.786760 quoi
1.769069 vision	1.777568 -
1.768280 honneur	1.758319 gens
1.763166 asie	1.747909 assez
1.762665 efficacité	1.741650 capables
1.745192 saluer	1.716491 penser
1.743871 soutien	1.700678 bref
1.737269 renforcer	1.688314 puisque
1.715155 concitoyens	1.672872 on
1.709736 réforme	1.662164 étais
1.703412 devons	1.620722 parle
1.695359 engagement	1.618184 fallait
1.689079 estime	1.604095 simplement
1.671255 titre	1.589586 entendu
1.669899 pleinement	1.580018 suite
1.662398 cœur	1.572140 peut-être
1.661476 ambition	1.571393 espère
1.654876 santé	1.560364 parlé
1.640298 stabilité	1.550856 dis
1.632421 amitié	1.549594 cela
1.628630 accueil	1.538523 existe
1.622473 publics	1.535598 façon
1.616558 diversité	1.529225 pourrait
1.614945 service	1.525645 là
1.612488 valeurs	1.525508 chose
1.610123 détermination	1.523575 époque
1.601097 réformes	1.522290 production
1.592938 état	1.519365 trouve
……..	……..

Classification with FMC

Dickens-Collins contreversy (stylometry)

DICKENS	
1,227361	coming
1,04304	heart
1,197376	going
1,531862	boy
1,073734	hands
1,379369	cried
1,10153	men
1,491942	gentleman
1,261003	street
1,316684	dear
1,022143	love
1,175788	like
1,240113	young
1,04989	light
1,151491	seemed
1,194106	dark
1,16963	happy
1,130386	know
1,584914	indeed
1,507332	fire
1,50468	often
1,513448	great
1,455893	pretty

DICKENS CHILDNESS (IDF only)

1,480477	delighted
1,477346	laugh
1,480015	observed
1,782888	boots
1,638954	blessed
1,146102	walk
1,57313	piece
1,585134	played
1,562992	rolling
1,49908	sing
1,463912	horses
1,439286	worked
1,436957	sun
1,4481	comfortable
1,454421	touching
1,389128	teach
1,461329	pleasant
1,531391	shadows
1,476144	windows
1,396892	pains
1,298137	youre
1,405204	raising
1,238744	wall

COLLINS	
1,294836	speaking
1,534484	answered
1,546532	waiting
1,250605	heard
1,575393	servant
1,232137	interest
1,064008	woman
1,288278	led
1,332932	left
1,299098	feel
1,123745	remember
1,069823	met
1,246964	open
1,321961	will
1,257266	look
1,064291	means
1,419852	husband
1,390256	doctor
1,538165	present
1,193449	suddenly
1,348551	herself
1,376969	truth
1,174412	letter

COLLINS CHILDNESS (IDF only)

1,454853	proceeding
1,341769	explain
1,459178	anxiety
1,374366	noticed
1,567781	discovery
1,519614	decide
1,486814	events
1,542078	informed
1,53803	approached
1,449153	eagerly
1,709461	confession
1,429975	accepted
1,450273	necessity
1,464363	evidence
1,416902	address
1,41972	capable
1,464345	patience
1,540817	sadly
1,386233	entering
1,416557	importance
1,685569	resolution
1,364304	alarm
1,45935	possessed

The method provide exhaustive and precise results and allows fine-grained analysis modulation.

Clustering with FMC

Structural pattern of French verbs [Falk – ACL 2012, Lamirel 2014]

IGNGF clustering is a parameter-free incremental neural clustering method exploiting feature maximization in substitution to standard distances (Euclidean, cosine, …)

- Combines clustering and feature selection/explanation capabilities (pseudo-symbolic behavior)
- Shown to outperform both state-of-the art symbolic (FCA) and numeric (Spectral Clustering) methods in complex problems: clustering of French verbs using syntactic-semantic features [Falk 12]
- Can be used to highlight latent classes in classification problems: classification of institutional websites using communication signatures [Lamirel 13]

Discriminant and shared characteristics
F-measure > F-average

Marginal characteristics
F-measure < F-average

Verbs ranked by representativeness / class

C6- 14(14) [197(197)]

Prevalent Label — = AgExp-Cause

0.341100 G-AgExp-Cause
0.274864 C-SUJ:Ssub,OBJ:NP
0.061313 C-SUJ:Ssub
0.042544 C-SUJ:NP,DEOBJ:Ssub
****
****
0.017787 C-SUJ:NP,DEOBJ:VPinf
0.008108 C-SUJ:VPinf,AOBJ:PP
...

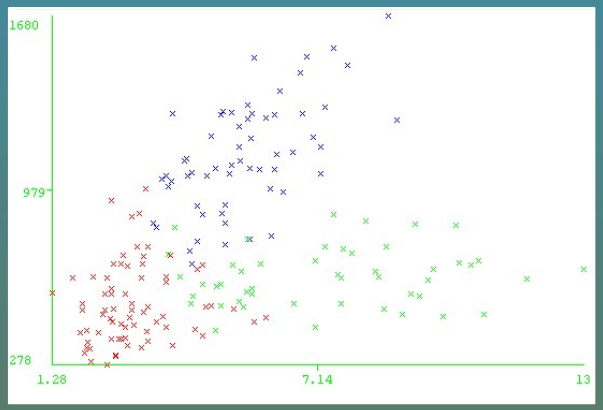
[**déprimer 0.934345 4(0)] [affliger 0.879122 3(0)] [éblouir 0.879122 3(0)] [choquer 0.879122 3(0)] [décevoir 0.879122 3(0)] [décontenancer 0.879122 3(0)] [décontracter 0.879122 3(0)] [désillusionner 0.879122 3(0)] [**ennuyer 0.879122 3(0)] [fasciner 0.879122 3(0)] [**heurter 0.879122 3(0)] . . .

F-max cluster labeling can be exploited with any clustering method.

A simple numerical case

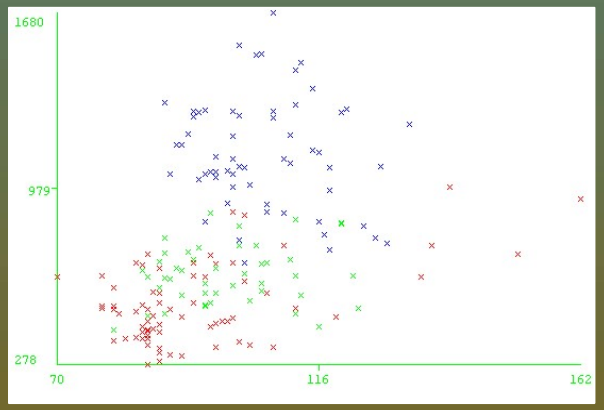
Wine dataset with J48 or FMC

J48

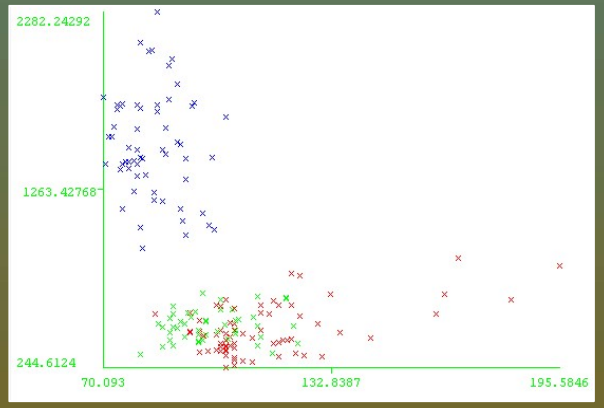


J48 and FMC select both 2 features among 13 but discrimination become better with FMC when magnification factor is increased

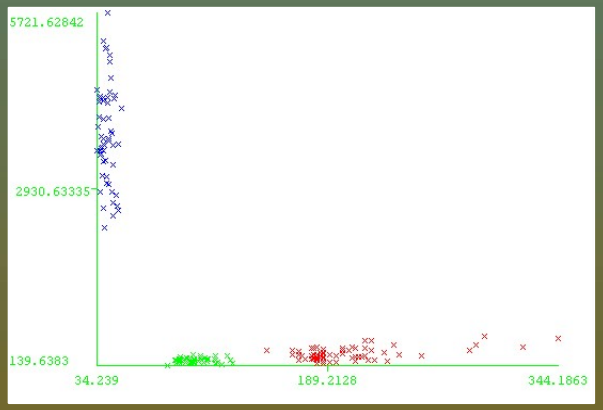
FMC



k = 1



k = 2



k = 4

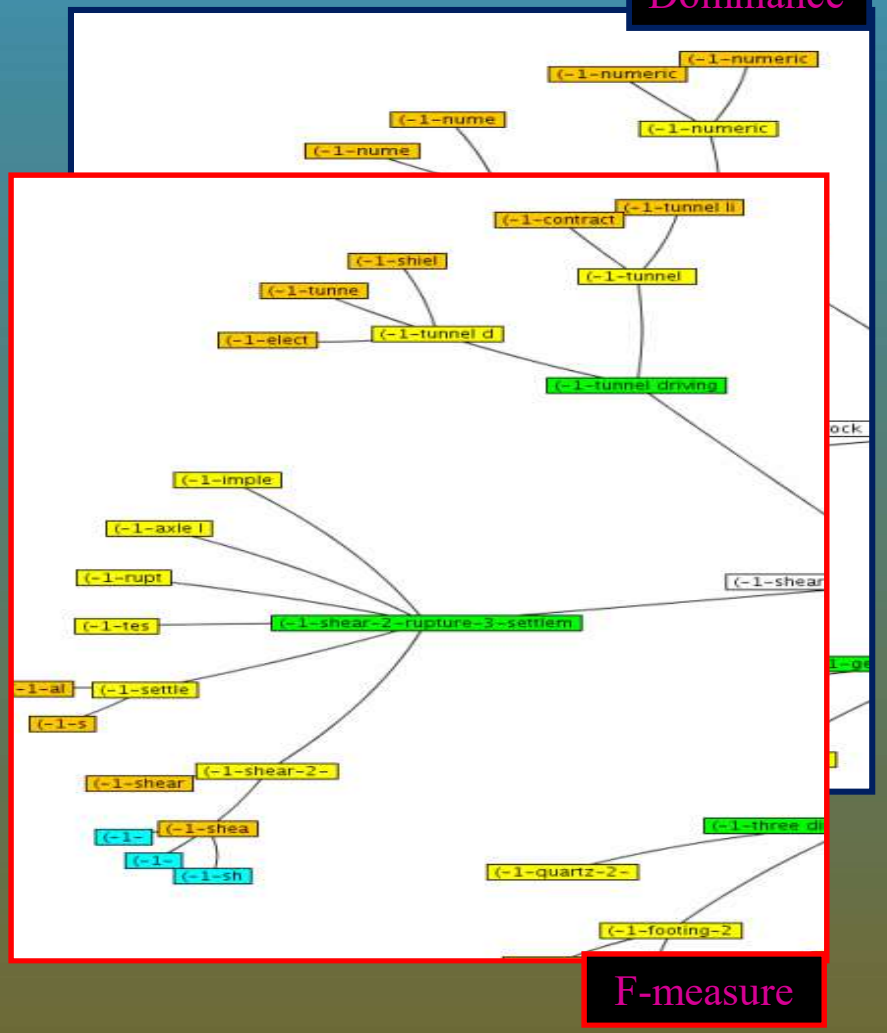
Cluster labeling based on feature maximization

Labeling method comparison [Lamirel 2008]

	PLS	AVP	PSS	PHL
Dominance	1216	0.03	568	525
Frequency	245	0.24	166	592
F-Measure	155	0.26	112	760
Chi²	121	0.21	89	1485

PLS : Penalty of Leave Similarity,
ALP : Average Leave labels Precision,
PSS : Penalty of Sons Similarity,
PHL : Penalty of Labeling Heterogeneity.

Dominance



F-measure

F-measure provides the best compromise: exhaustivity – discriminance [Lamirel 08].

Contrast graphs based on F-max metric

Contrast graphs

Principle

- Contrast graphs are bipartite graphs based on the relations between a set of features S and a set of labels L [Cuxac 2013].
- Theoretically, the set of labels L could represent any kind of information to which features can be related with and the set of features S is a subset of a global feature set F (i.e. the original feature space on which rely the data of a dataset) that has been obtained through a feature selection process, like feature maximization.
- In the case of the use of feature maximization, the weight c(u,v) of an edge (u,v), $u \in S$, $v \in L$ represents the contrast of feature u for a label v.

Contrast graphs

Principle

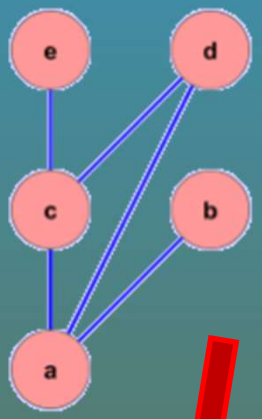
Such kind of graphs have many interesting properties :

- They reduce the cognitive overload produced with classical graphs representation because of the associated feature selection process
- They can be used to indirectly highlight relationships between labels, whenever features have contrasted interaction with several labels.
- Third, the combination of this approach with weighted force-directed model for graph representation permits altogether to highlight central or most influent labels of the L set and to easily identify the labels that are the most densely connected through associated features.

Contrast graphs

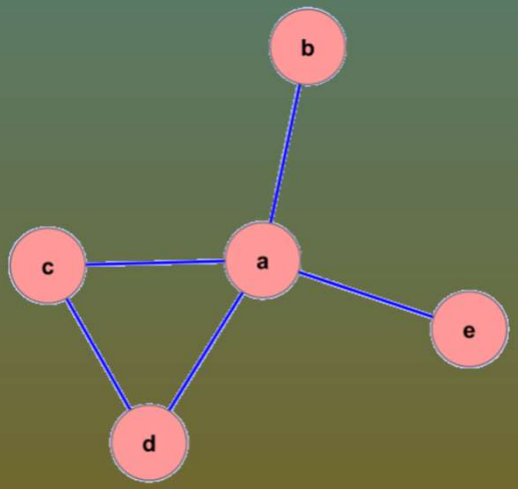
Principle

Node S	Node T	Edge Weight
a	b	1
a	c	1/4
a	d	1/4
a	e	1/10
c	d	1/4

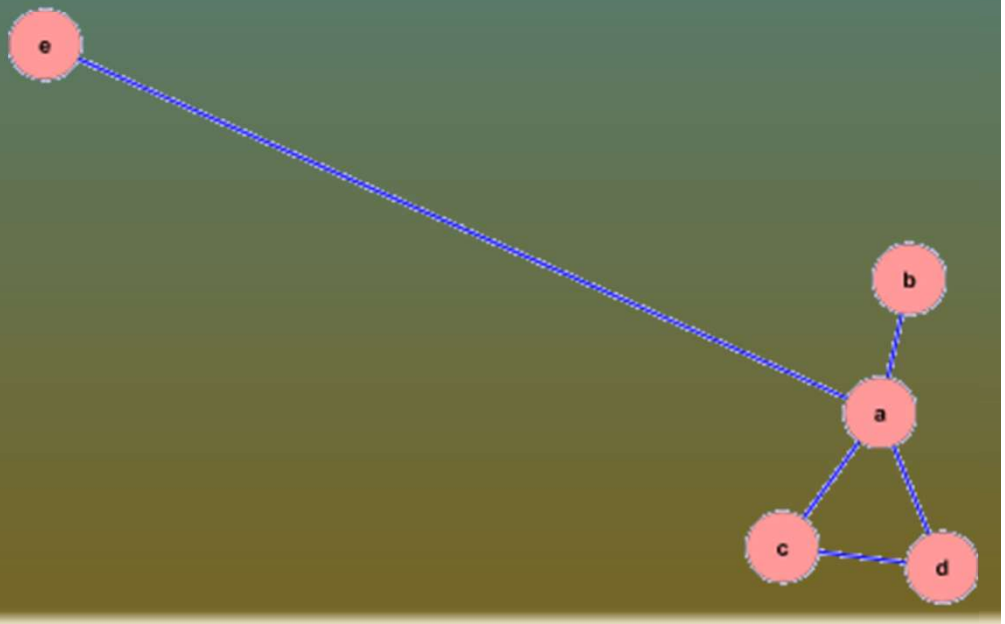


Contrast issued from feature maximization metric is used for materializing force.

Force directed graph

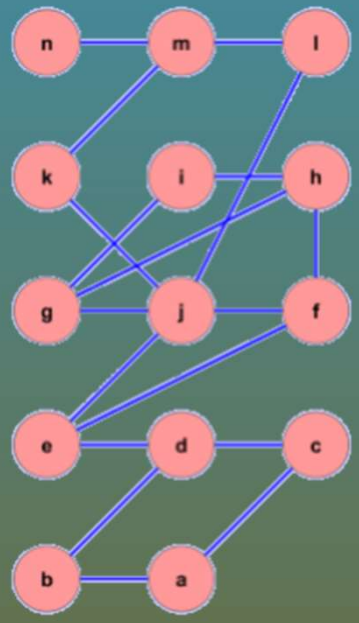


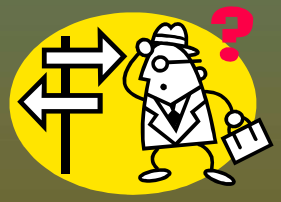
Force directed weighted graph

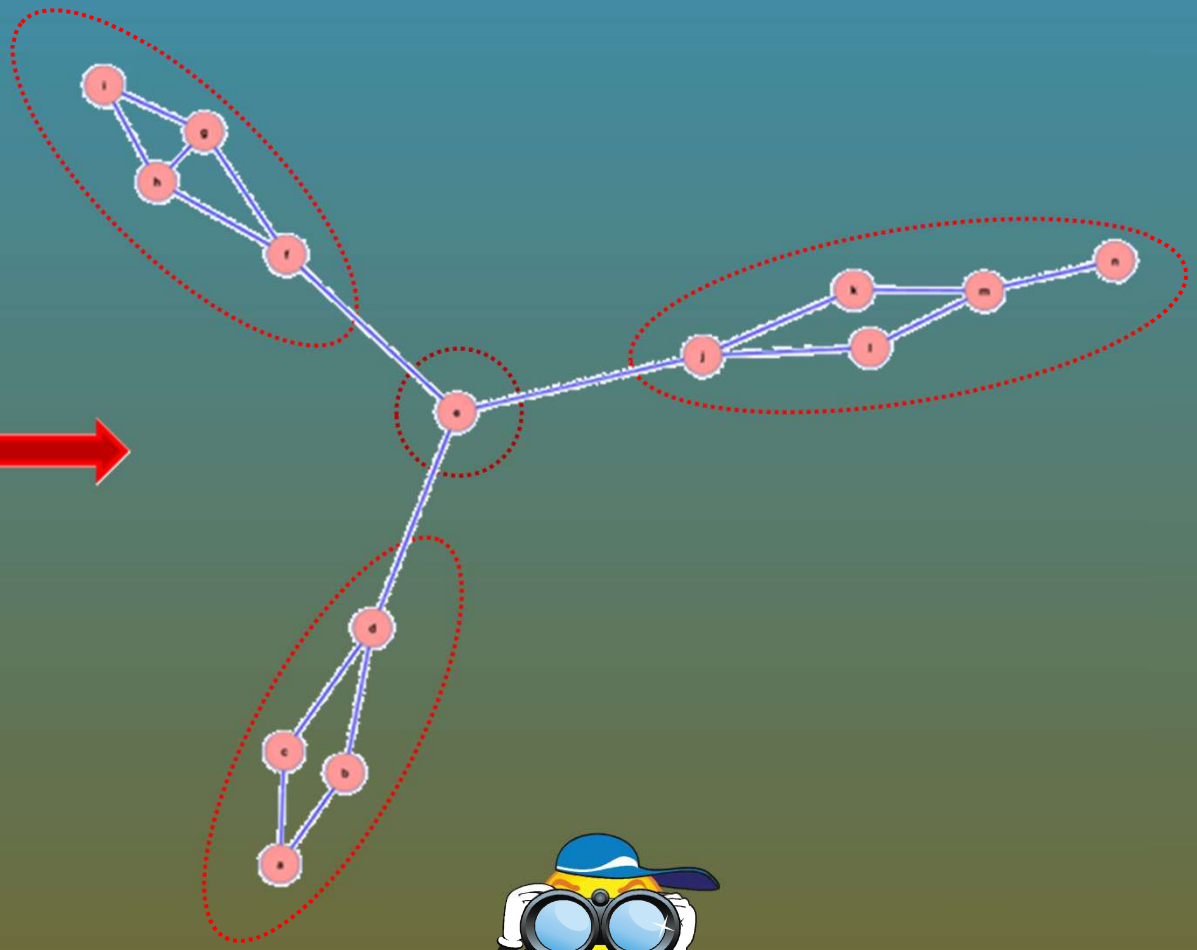


Contrast graphs

Principle

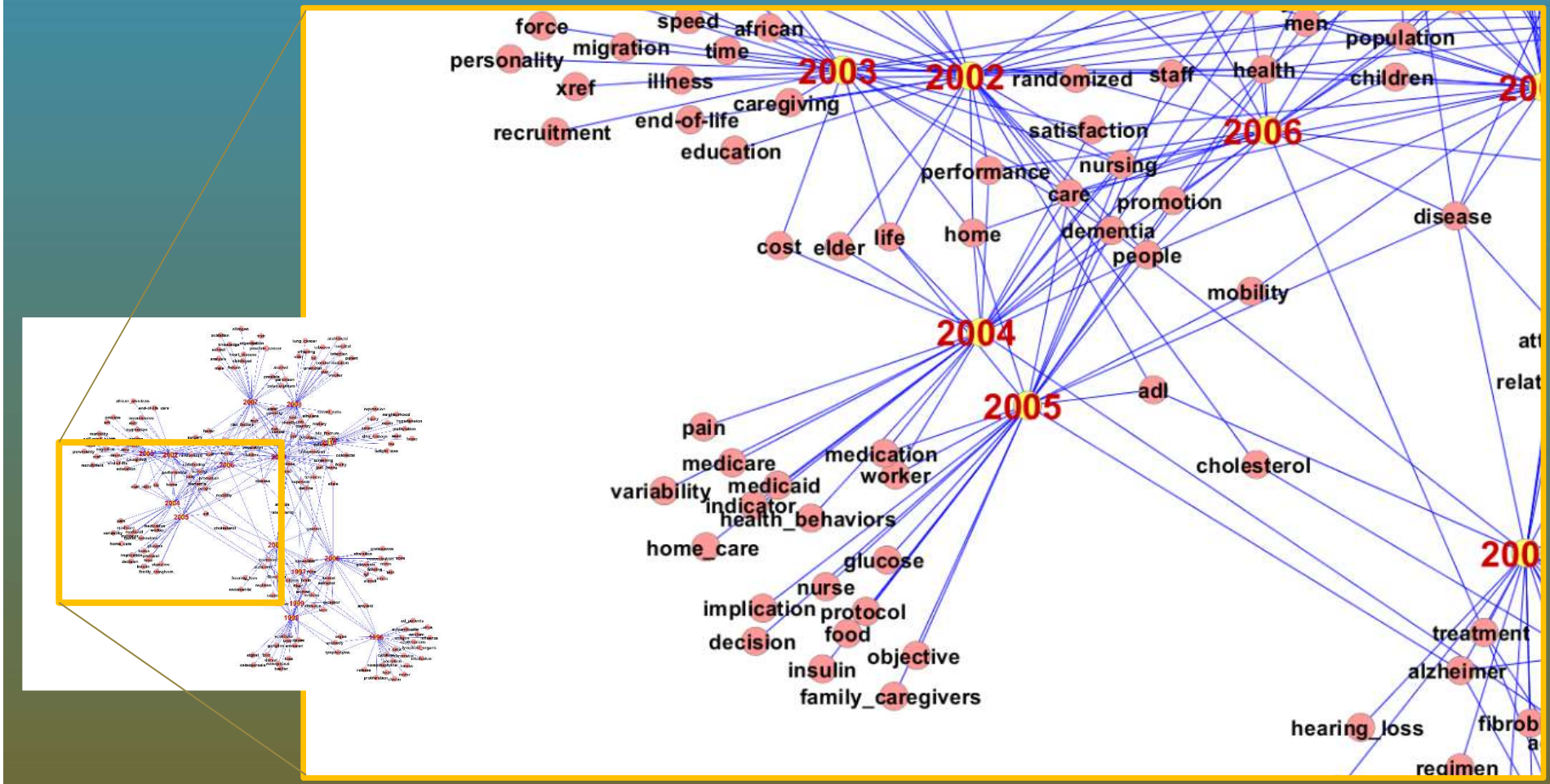






Contrast graphs with classified data

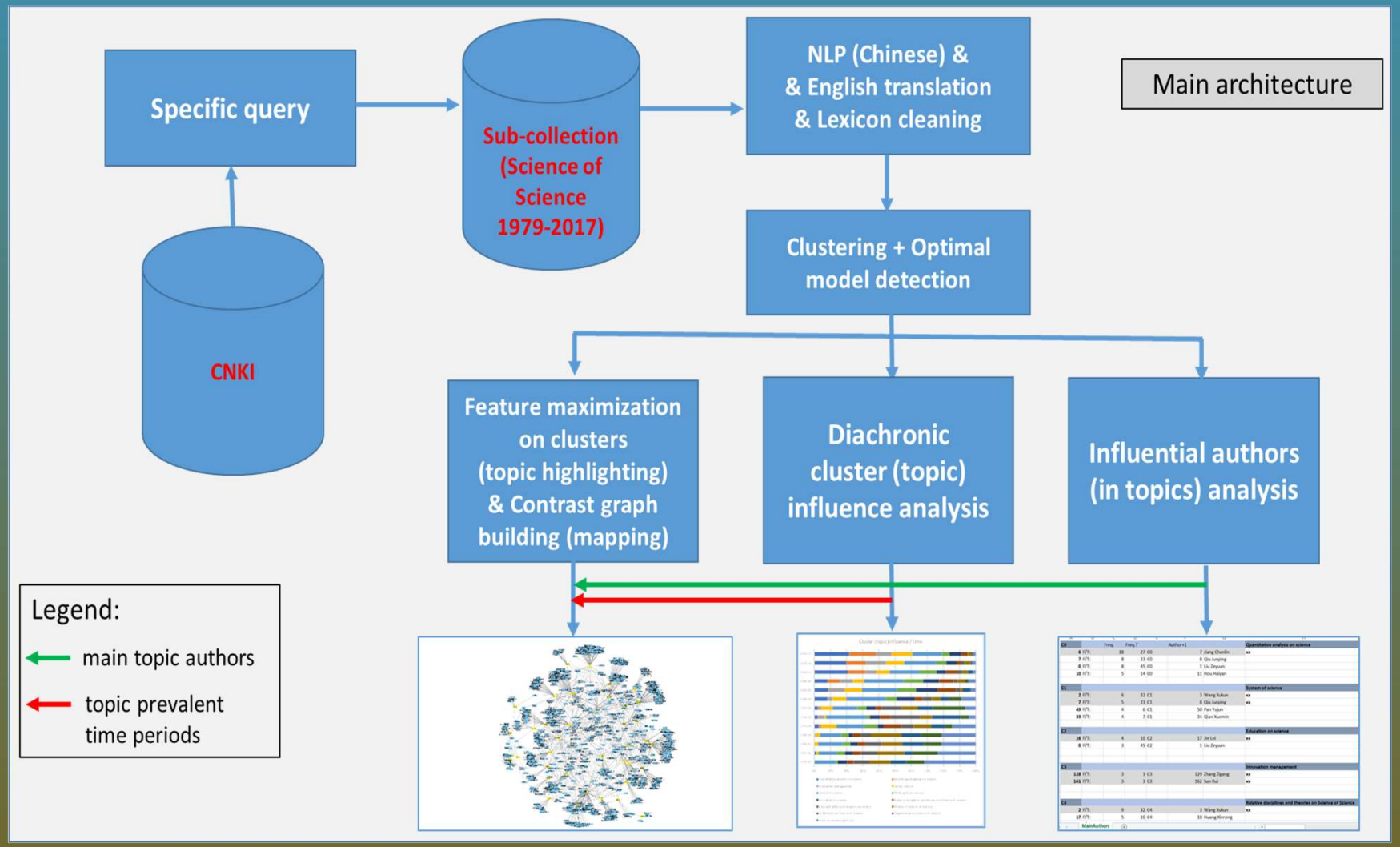
Health care data in years



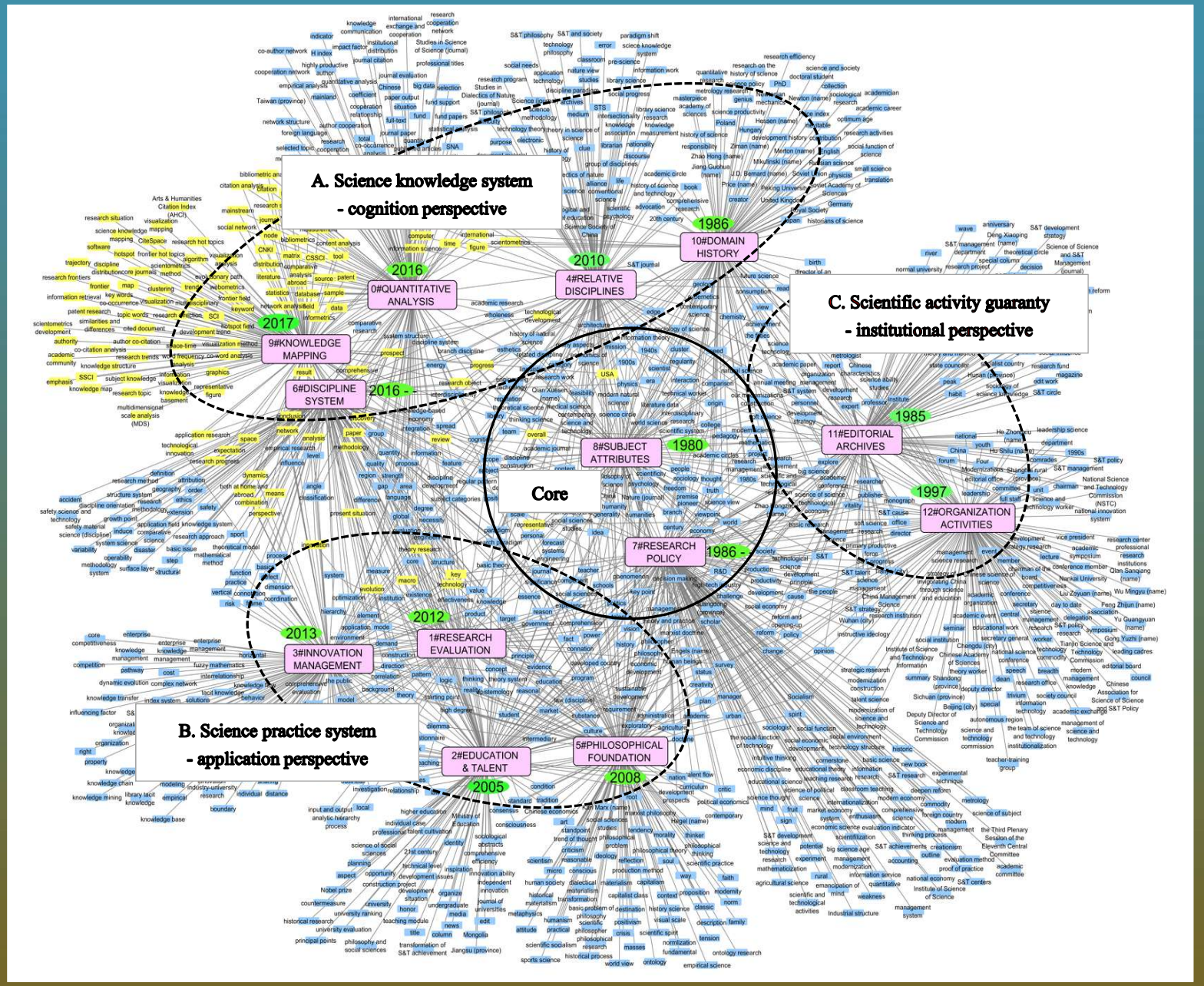
Graphs identify essential relations between categories and features also highlighting highly central features.

Contrast graphs with classified data

Map of Science of Science history in China [Lamirel 2020]



Science of Science in China over 40 years [Scientometrics 2020]



A new reliable approach for community role detection using an adaptation of F-max metric [Dugué & Lamirel 2018]

Community roles

- Many real-life system are designed in the form of networks,
- Some current examples are social networks, biological networks, road and transportation networks, collaboration networks, lexical networks,
- Such networks are important data sources for understanding the system they model,
- From early years, some important notion like centrality have been studied for many purposes (leadership detection, information spreading, organizational efficiency ….).

Centrality measures

Considering that a network can be defined as a graph $G(V,E)$ with V being a set of nodes and G being a set of edges between nodes.

Betweenness centrality [Freeman 79] is defined as :

$$C_b(u) = \sum_{v,w \in V, v<w} \frac{\sigma_{vw}(u)}{\sigma_{vw}}$$

where $\sigma_{vw}(u)$ is the number of paths between nodes v and w that pass through the node u and σ_{vw} is the total number of paths between nodes v and w.

Centrality measures

Proximity centrality can be defined as :

$$C_c(u) = \frac{1}{\sum_{v \in V} dist(u,v)}$$

where dist (u,v) is the geodesic distance between nodes u and v.

PageRank [Brin et Page 2012] is defined for an oriented network as :

$$PR(u) = (1-d) + d \sum_{v \in N_u^-} \frac{PR(v)}{|N_v^+|}$$

where N_u^+ and N_u^- represent respectively the entering and the exiting neighbors of a node u.

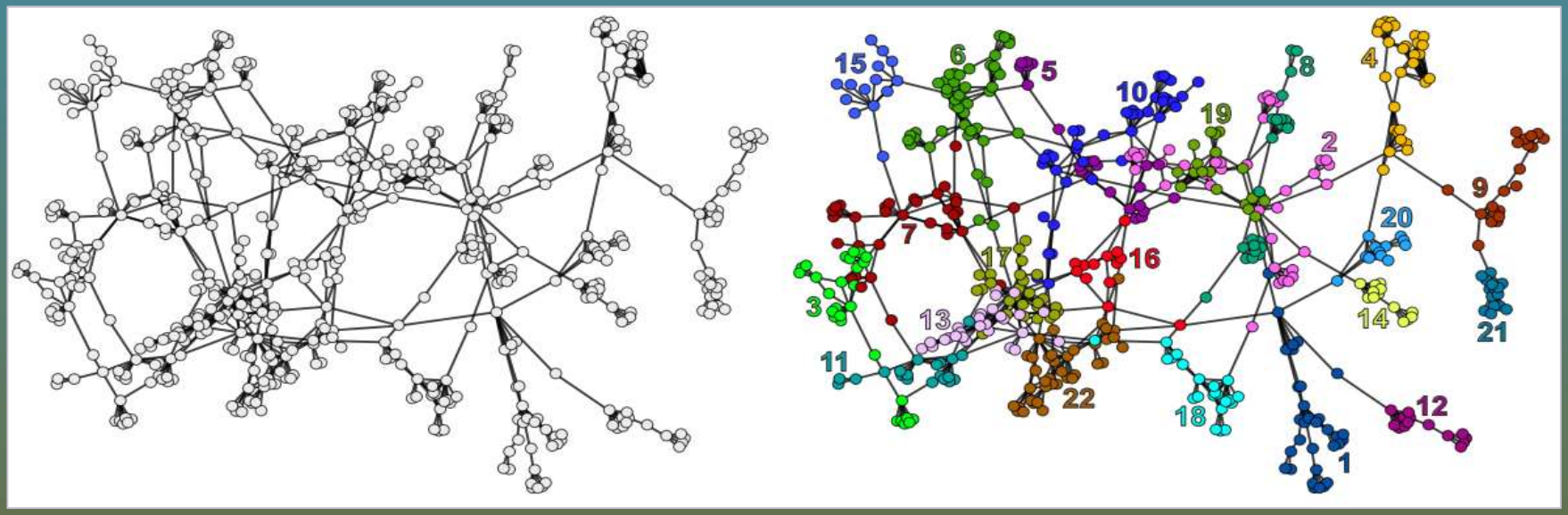
Community role measures

Centrality measures do not take into account a more recent reality related to the big data context with networks of growing size :

- Most of the big networks have a community structure (small-word phenomenon) [Newman et Girvan 04],
- The networks can be thus studied at a mesoscopic level instead to stay at a global level,
- New measures have been proposed for that purpose.

Community role measures

Community structure



This figure shows the community structure (right) after community detection being applied on the network (left).

Community role measures

Guimera et Amaral (2005) propose two different measures for community role detection :

Intra-module Degree :

$$Z_i(u) = \frac{d_i(u) - \mu_i(d_i)}{\sigma_i(d_i)}$$

Participation Coefficient :

$$P(u) = 1 - \sum_i \left(\frac{d_i(u)}{d(u)}\right)^2$$

where $d_i(u)$ is the degree i.e. the number of edges of the node u in the community i, $d(u)$ its total number of edges, $\mu_i(d_i)$ is the average degree of the node in the community i and $\sigma_i(d_i)$ the related variance.

Community role measures

Guimera and Amaral (2005) defined different types of communities and node roles depending on the values of Intra-module Degree and Participation Coefficient :

Intra-module degree		Participation coefficient	
Hub	>= 2.5	Provinvial Connector Ultra-Connector	<= 0.30]0.30 ; 0.75] > 0.75
Non-hub	< 2.5	Ultra-Peripheric Peripheric Connector Ultra-Connector	<= 0.30]0.05 ; 0.62]]0.62 ; 0.80] > 0.8

Community role measures

Guimera and Amaral approach causes however some problems to be generalized outside of the scope of biological networks [Dugué 15] :

- Thresholds defined for the different categories of nodes are not universal and thus difficult to estimate in most cases,
- Values of participation Coefficient on networks where highly connected nodes have a lot of external connections could be abnormally low because the measure encapsulates several aspects of node connectivity.

New adapted CR measures based on Feature Maximization

Node Recall :

$$NR_i(u) = \frac{d_i(u)}{d(u)}$$

Node Predominance :

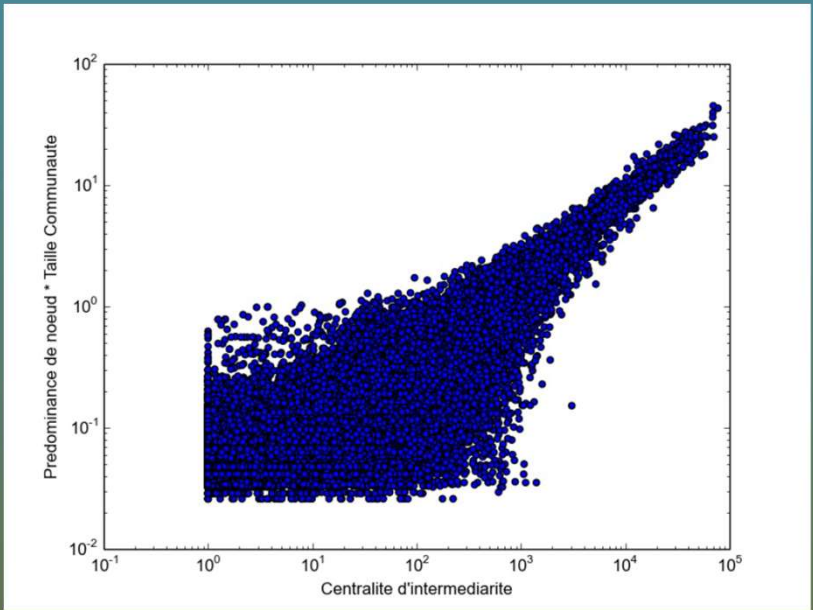
$$NP_i(u) = \frac{d_i(u)}{d_{c_i}}$$

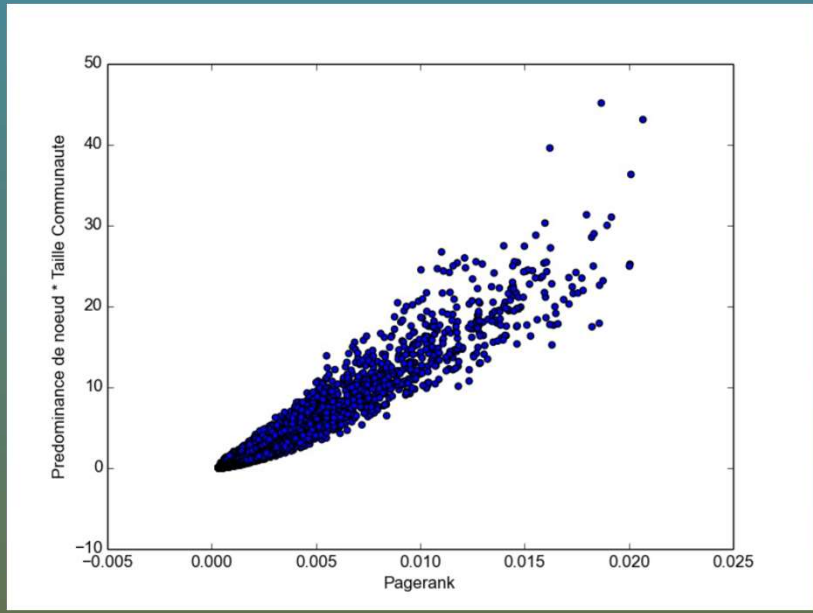
where $d(u)$ is the degree (i.e. sum of the links) of node u, $d_i(u)$ the degree of node u in the community i and d_{c_i} represents the sum of the links in the community i.

New adapted CR measures based on Feature Maximization

- Node Predominance is used to characterize the embedment of a node in its community. It is similar to the embedment measure proposed by Lancichinetti et al. (2010),
- Node Recall is useful to highlight the connectivity of a node with the nodes which are external to its community. A weak node Recall means that a node is more connected outside its community,
- Provincial hubs can be easily detected with the help of this two measures, but analysis can be extended to other kinds of hubs and non-hubs sub-categories.

Correlation with centrality measures





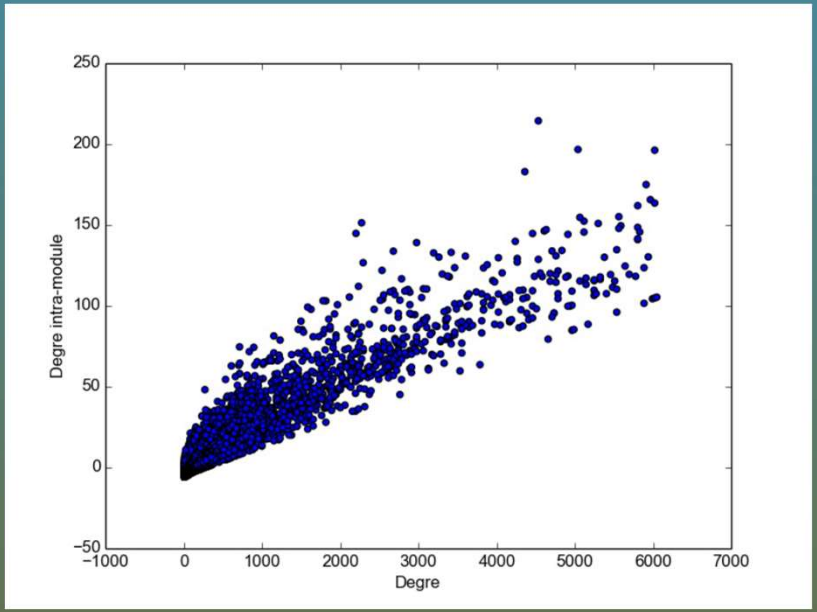
Node Predominance vs Betweenness Centrality (at a log scale) and Node Predominance vs Pagerank : high correlations exist.

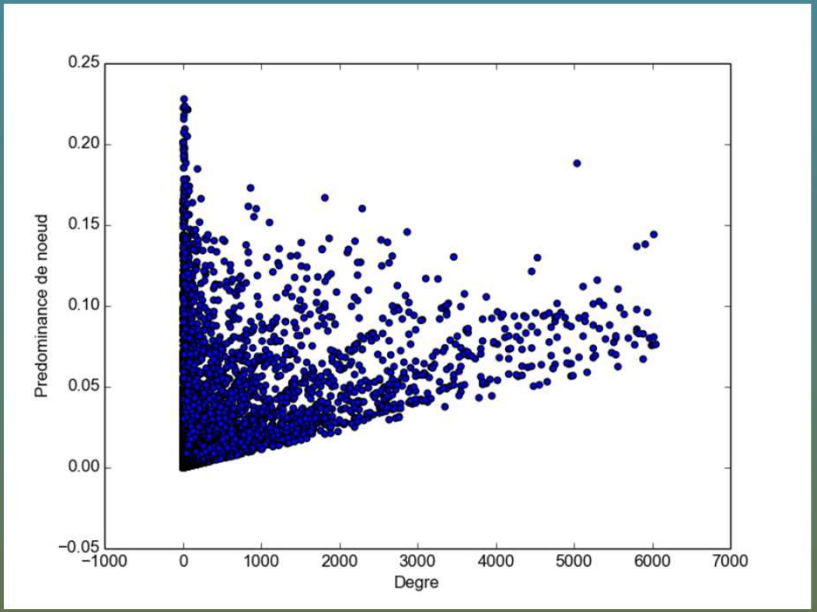
Node Predominance is multiplied by the size of communities in all cases.

Correlation with centrality measures

- Independence between Node Predominance and Proximity Centrality can be observed,
- Our measure is also independent of the maximal value of *k-core* of the community to which the nodes belong. Similar observation can be done for Intra-module Degree,
- A *k-core* is a subgraph a given graph in which the degree of the associated nodes is > *k*. This measure is important for studying propagation of infection in a graph.

Correlation with measures associated to community roles

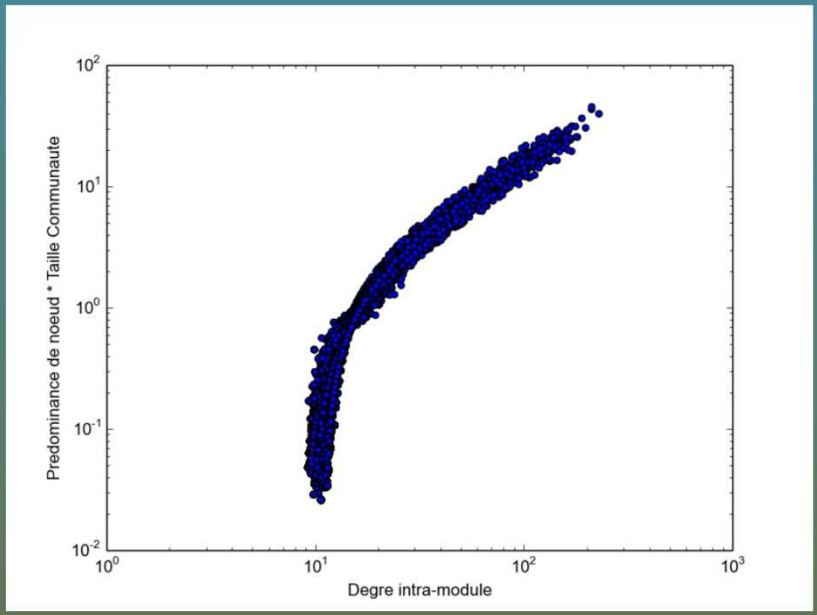




Node Predominance vs Intra-module degree depending on the average degree of the nodes in the communities :

Node Predominance is almost independent of the degree of the nodes.

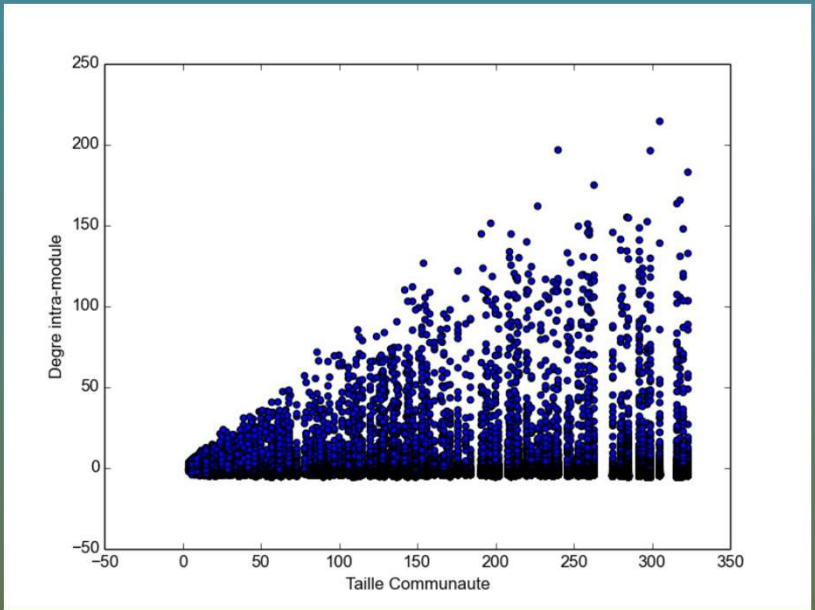
Correlation with measures associated to community roles

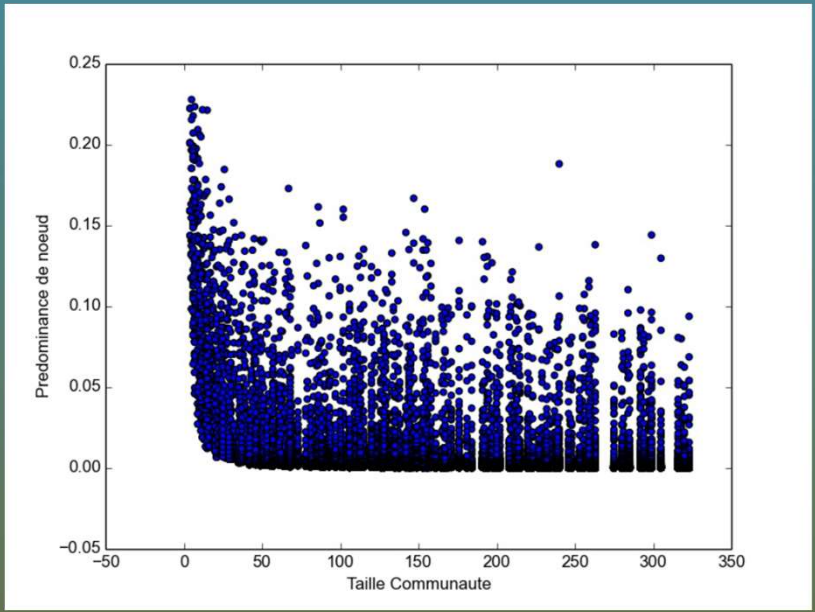


Node Predominance multiplied by the size of the communities
vs Intra-module Degree :

Measures are highly correlated but Node Predominance is independent of the size of the communities.

Correlation with measures associated to community roles

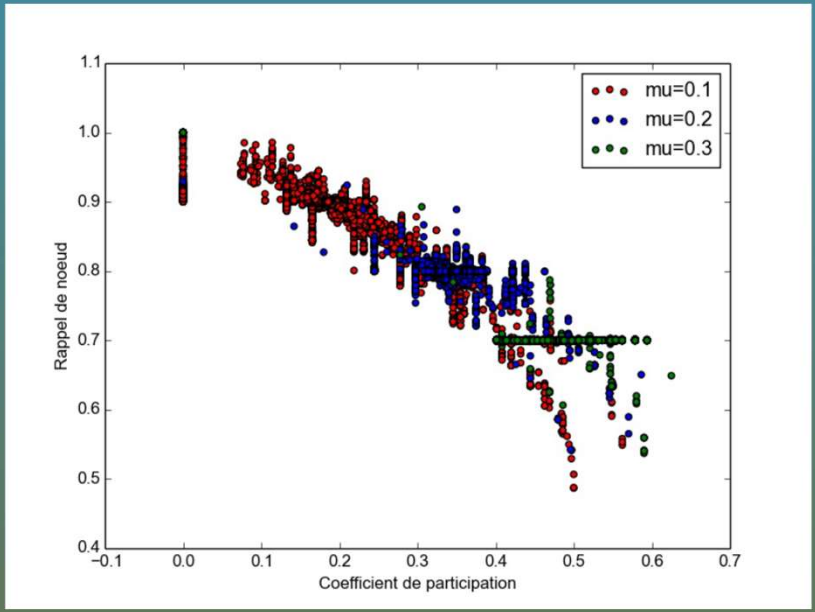




Node Predominance vs Intra-module Degree depending of size of the communities :

Node Predominance is almost independent of the size of the communities.

Correlation with measures associated to community roles



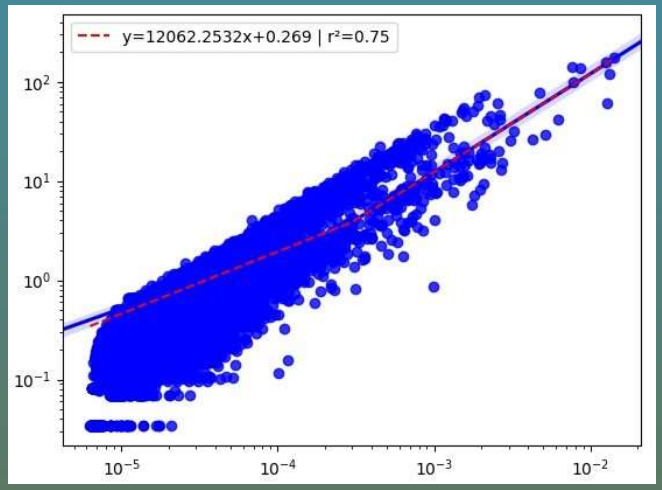
Node Recall multiplied by the size of the communities vs Participation Coefficient :

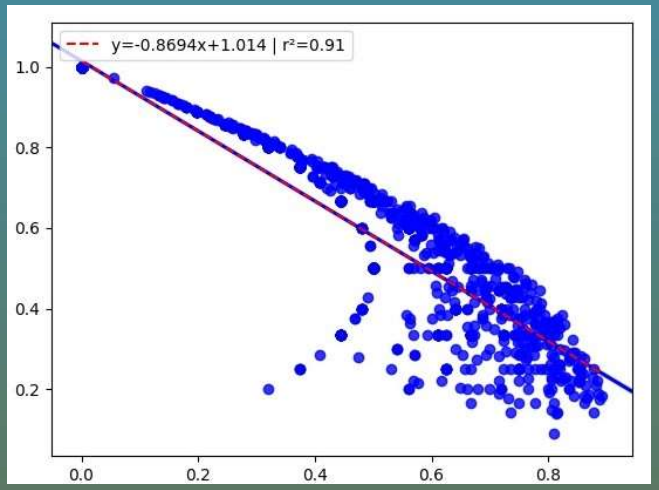
High correlation can be observed but Participation Coefficient is weakening as compared to Node Recall for high values of *mu*.

=> Participation Coefficient tends to be biased or abnormally weak when high degrees node in one community are highly connected with nodes that are external to said community.

Correlation with other measures

On real graphs





Node Predominance x c. size
vs PageRank

Node Recall x c. size
vs Participation Coefficient

Same observations hold but even more precise observations can be made on real graphs.

Community role detection with F-max metric

- The approach has many advantages as compared to usual approaches because it is a parameter-free approach that is independent of the size of the communities and of degree of the nodes,
- The computation time of the proposed measures is low and it can thus be exploited in a big data context,
- More exhaustive tests must be conducted on real-life networks to confirm the very promising properties of this measure,
- Because of its advantages the method is prone to the analysis of the evolution of communities during time (see further).

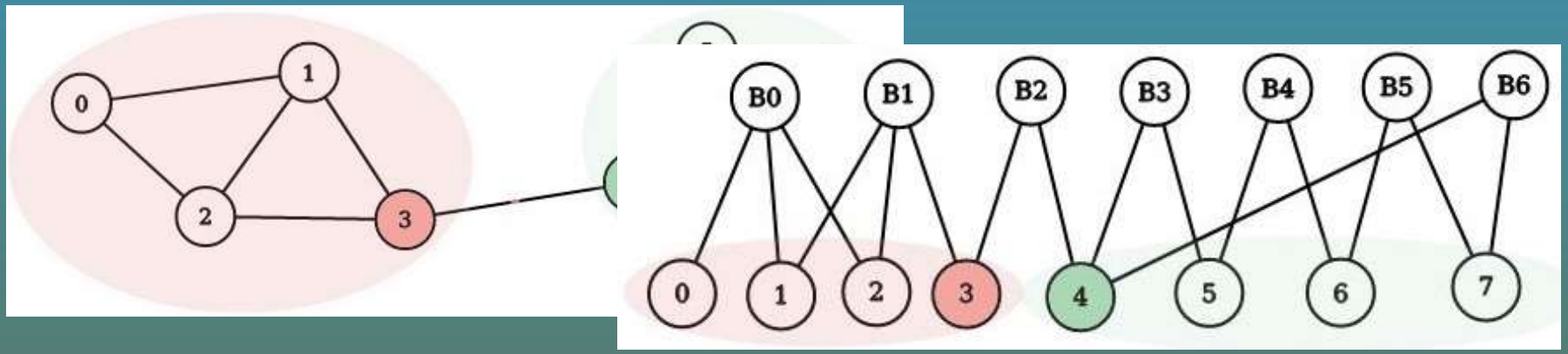
Sparse node embedding in graphs

Principles [Prouteau et al. 2021]

- Each initial unipartite graph constituted of B nodes is transformed into a bipartite graph with a minimal set of T nodes connected to the initial set of B nodes,
- For that purpose, the principle to find an approximate solution of the graph minimum clique covering problem [Guillaume et al. 06], by the use of a community detection algorithm,
- The embedding space is the represented by obtained communities,
- Node embedding is performed by the use of node recall and node predominance measures,
- It results in a sparse interpretable representation,
- Computation cost of the whole process is near linear.

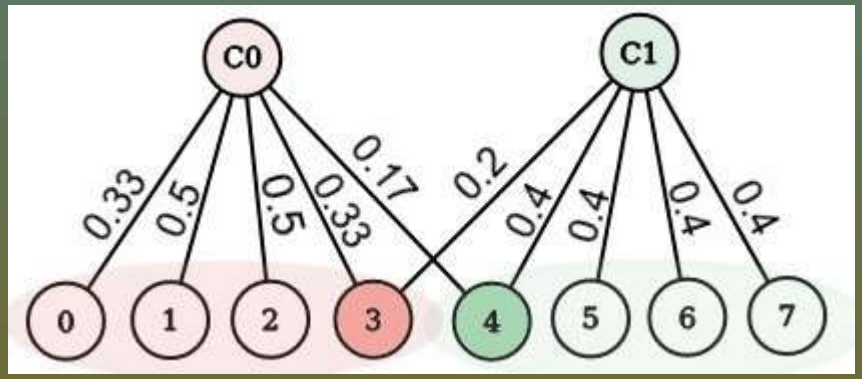
Sparse node embedding in graphs

Example



Initial unipartite graph

Latent embedding structure with maximum cliques



Community detection and node predominance values

3	0.33	0.66	0.2	0.33
4	0.17	0.33	0.4	0.66
7	0	0	0.4	1

Resulting embedding vector

Sparse node embedding in graphs

Results (link prediction)

	SINr	SINr-SVD	Heuristics	DW	WL	HOPE
Cora	**0.83**	**0.83**	0.76	0.73	0.82	0.75
Eu	**0.88**	**0.88**	0.87	0.81	0.87	0.87
Cts	**0.88**	0.86	0.78	0.76	0.87	0.83
arXiv	0.92	0.90	**0.97**	0.92	0.96	0.92
Fb	0.91	0.89	**0.93**	0.86	0.92	0.90

Comparison with state of the art methods

	Cora	Eu	Cts	arXiv	Fb
SINr	0.3/1.3	0.3/2	0.1/0.9	0.9/4	3/8
HOPE	0.2/3	0.6/8	0.7/2	10/120	26/195
DW	24/36	13/18	20/30	264/378	336/422
WL	26/38	12/18	24/36	261/365	475/652

Runtime and CPU time

Whenever community structure is sufficiently present, method outperforms usual network embedding methods with low cost and high resulting sparsity. Performances are preserved when SVD-based space compression is performed.

	p	Q_c	σ
Cora	23	0.80	0.94
Eu	6	0.41	0.44
Cts	33	0.85	0.96
arXiv	28	0.62	0.87
Fb	73	0.62	0.96

Number of communities, modularity, sparsity score

Sparse word embedding

Principles [Prouteau et al. 2021]

- The graph of word co-occurrences is computed using a sliding window on texts,
- The obtained cooccurrence graph is filtered by the use of point-wise mutual information test (~independence test),
- The edges in the graph are ranked by decreasing value the sum of degrees of their related nodes and are iteratively reweighted using:

$$IPMI(u,v) = \frac{W_{u,v}}{d(u)d(v)}$$

where $W_{u,v}$ is the cooccurrence count between u and v, d(x) is the degree of node x,

- Similar process than the one presented for general graph (community, detection and community-based embedding) is used for generating word embeddings.

Sparse node embedding in graphs

Results (link prediction)

text8	W2V	GloVe	SVD2vec	SINr
MC28	0.58	0.42	0.67	**0.69**
RG65	0.52	0.48	0.57	**0.64**
MTurk771	**0.52**	0.48	0.45	0.48
WS353Rel	0.47	0.44	0.46	**0.50**
WS353Full	**0.55**	0.47	**0.55**	0.53
MEN	0.53	0.48	**0.61**	0.52

OANC	W2V	GloVe	SVD2vec	SINr
MC28	0.45	0.54	0.33	**0.62**
RG65	0.33	0.32	0.32	**0.39**
MTurk771	**0.44**	0.39	0.36	0.37
WS353Rel	0.40	0.34	**0.47**	0.41
WS353Full	0.49	0.40	**0.51**	0.44
MEN	0.44	0.46	**0.59**	0.40

Comparison with state of the art word embedding methods
(Spearman similarity with expert evaluated word similarity)

The method achieves the best performance with pair of words of the same type (i.e., pair of names) whilst reaching sparser representation in embedding space and achieving much faster computation time.

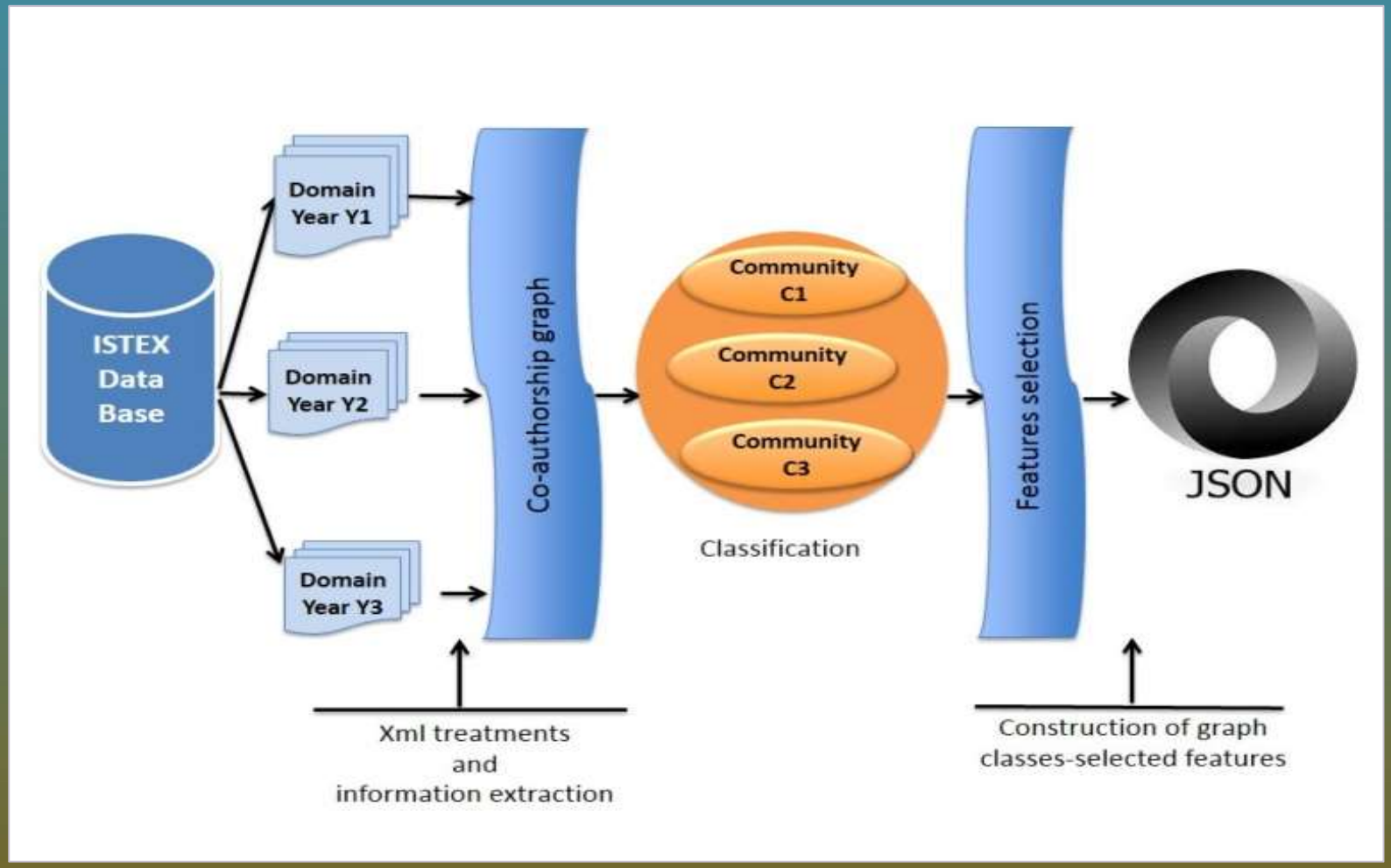
Diachronic community analysis

The CGEM (Collaboration Graph Evolution Monitoring) method [Dugué & Lamirel 2017]

- An ISTEX database is queried to produce an initial corpus,
- The documents are split into sub corpora that represent different publishing periods,
- Weighted undirected collaboration graphs of each period,
- Community detection is made using the INFOMAP algorithm;
- Salient authors are extracted for each community of each period using new measures of community role detection,
- ….

Diachronic community analysis

The CGEM (Collaboration Graph Evolution Monitoring) method

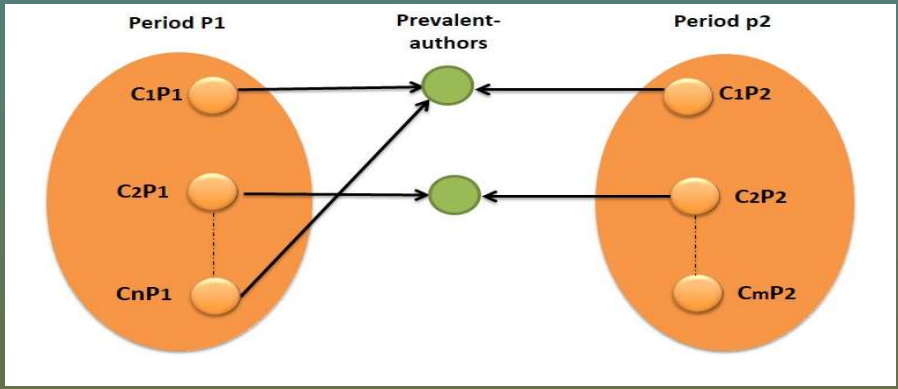


CGEM architecture

Diachronic community analysis

The CGEM (Collaboration Graph Evolution Monitoring) method

- ….
- Diachronic analysis [Lamirel 2012] is applied to monitor community visualizations evolution between periods (JSON report and Gephi visualizations are generated in this step) :



Dataset of papers on Parkinson disease published between 2000 and 2010 and extracted form ISTEX database. Papers are separated in 2 periods (2000-2005, 2006-2010). 2538 distinct authors are found in first period and 5961 in the second. 709 authors are shared between periods.

Diachronic community analysis

MVDA Paradigm on salient nodes and time periods

Comparison is performed using an adaptation of MVDA Bayesian reasoning with :

$$P(t|s) = \frac{\sum_{l \in L_s \cap L_t} L - F(l)_t}{\sum_{l \in L_t} L - F(l)_t}$$

where L_x represent the set of salient nodes associated to the community x, and $L_x \cap L_y$ represent the common salient nodes, which can be called the **matching kernel** (i.e. common salient nodes), between the community x and the community y.

Diachronic community analysis

MVDA Paradigm on salient nodes and time periods

The **similarity** between a community s of the source period and a community t of the target period is established using :

- The average matching probabilities $P_A(x)$ of a period community
- The global average activity A_x generated by a period model on the model of the alternative period and its standard deviation σ_x

Similarity is found if :

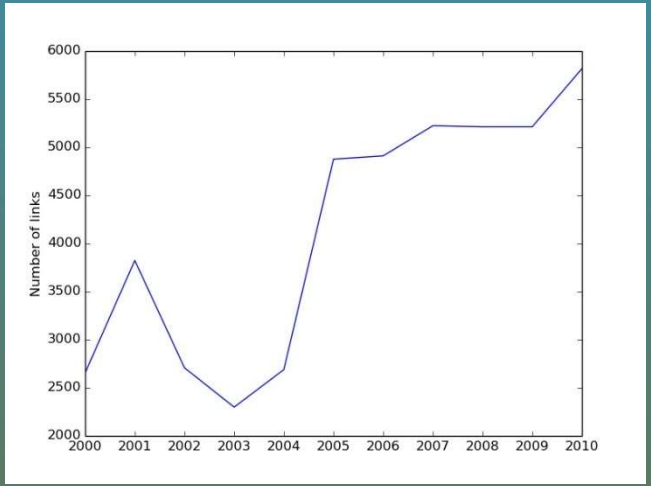
1) $P(t \mid s) > P_A(s)$ *et* $P(t \mid s) > A_s + \sigma_s$

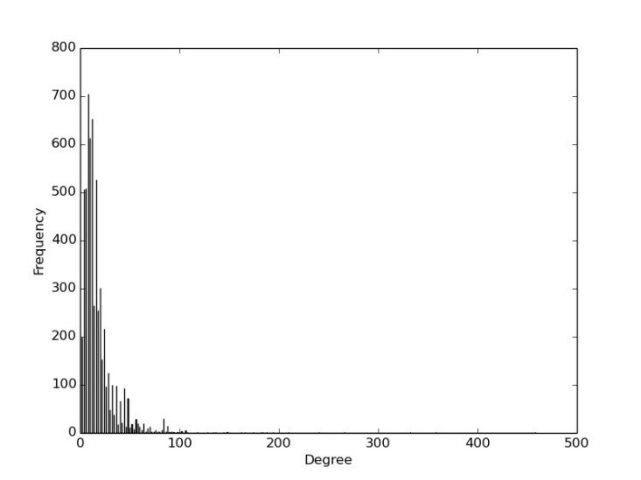
2) $P(s \mid t) > P_A(t)$ *et* $P(s \mid t) > A_t + \sigma_t$

Community splitting, community merging, vanishing communities, appearing communities events can be deduced from former similarity rules.

Diachronic community analysis

CGEM results - basic





Number of links of graphs obtained from the sub-corpora for each year period.

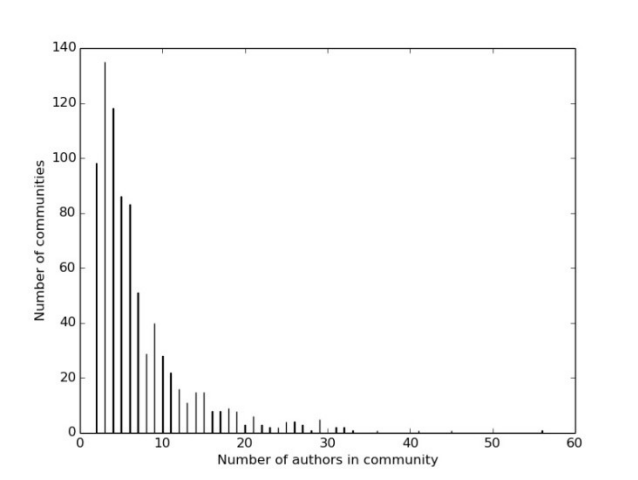
Value is globally increasing.

Degree distribution of graph from the second sub-corpora.

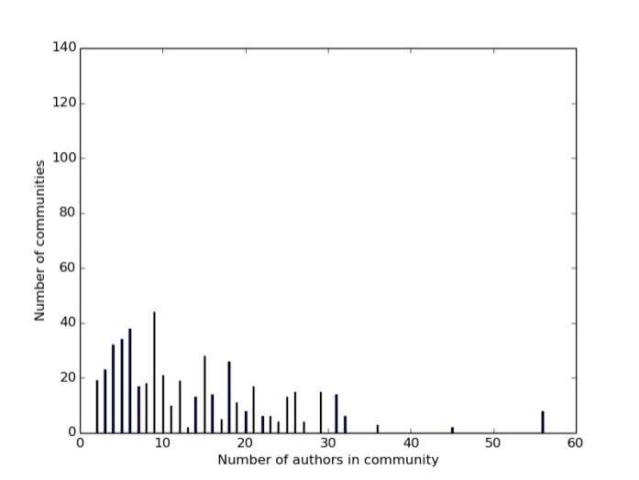
Follows a power law.

Diachronic community analysis

CGEM results – communities and matching kernels



Communities sizes distribution from the second sub-corpora.



Sizes distribution of the nodes matching kernel communities.

The communities constitute a high level description of the graph with sizes also following a power law.

The authors belonging to the communities matching kernels mostly belong to the biggest communities.

Diachronic community analysis

CGEM results – matching kernels description

Cluster Source	2
Cluster Target	48
Kernel Labels	

label	Chen Xiangmei
fSource	0.18604651
fTarget	0.15384616

label	Feng Zhe
fSource	0.10909091
fTarget	0.080000006

label	Fu Bo
fSource	0.062111802
fTarget	0.080000006

label	Hong Quan
fSource	0.18604651
fTarget	0.15384616

label	Shi Suozhu
fSource	0.10909091
fTarget	0.080000006

label	Wang Jianzhong
fSource	0.18604651
fTarget	0.15384616

label	Wu Di
fSource	0.062111802
fTarget	0.080000006

label	Xie Yuansheng
fSource	0.062111802
fTarget	0.080000006

Common Labels prevalent in Source

label	Lu Yang
fSource	0.16470589
fTarget	0.080000006

Common Labels prevalent in Target

label	Zhu Hanyu
fSource	0.025316456
fTarget	0.080000006

Matching reports highlight some temporal correspondence between communities belonging to different periods.

Common nodes in source and target communities (matching kernel)

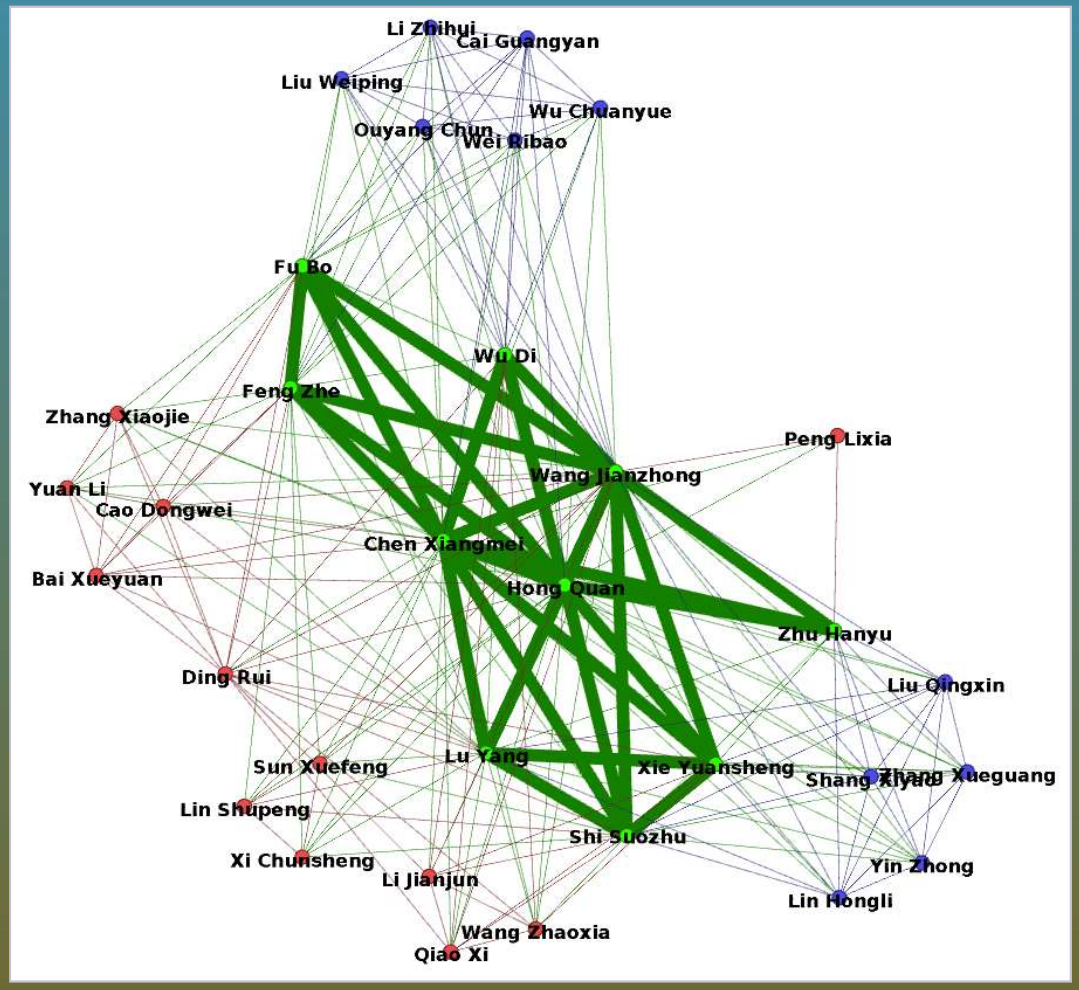
Specific nodes (i.e. context) of source community

Specific nodes (i.e. context) of target community

Inter periods communities matching report

Diachronic community analysis

CGEM results – matching kernels in networks



Inter periods communities matching graph

Among the 709 inter-period authors, 512 (i.e. 72%) of them are part of nodes matching kernel. It seems to indicate that salient authors in communities are particularly interesting to monitor. Indeed, they seem to form the **backbone** of knowledge production across time, linking different periods of time.

Synthesis

- We have presented two new approaches for graph-based data analysis based on Feature Maximization metric,
- Methods are complementary and prone to work in big data context with the main advantages to be being scale independent, parameter free and with high capacities of synthesis,
- These techniques proof to be especially useful for large scale science analysis,
- However, a richer domain of application is obviously emerging for such methods in the big data context (biological data, astronomical data, geographical data, ….).

References

External references are in the proposed papers

[Lamirel 2004] Lamirel J.-C., Al Shehabi S., François C., Hoffmann M.,
New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping,
Scientometrics, 60(3) 2004.

[Al Shehabi 2005] Al Shehabi S., Lamirel J.-C.,
Knowledge extraction from unsupervised multi-topographic neural network models,
Proceedings of ICANN 2005, Warsaw, Poland, September 2005.

[Attik 2006] Attik M., Al Shehabi S., Lamirel J.-C.,
Clustering quality measures for data samples with multiple labels,
Proceedings of IASTED International Conference on Databases and Applications (DBA), Innsbruck, Austria, February 2006.

[Lamirel 2008] Lamirel J.-C., Ta A.P., Attik M.,
Novel labeling strategies for hierarchical representation of multidimensional data analysis results,
Proceedings of IASTED International Conference on Artificial Intelligence and Applications (AIA), Innsbruck, Austria, February 2008.

References

External references are in the proposed papers

[Lamirel 2010a] Lamirel J.-C., Ghribi M., Cuxac P.,
Unsupervised recall and precision measures: a step towards new efficient clustering quality indexes,
Proceedings of 19th International Conference on Computational Statistics (COMPSTAT'2010), Paris, France, August 2010.

[Lamirel 2010b] Lamirel J.-C., Priyankar N., Cuxac P., Safi G.,
Mining research topics evolving over time using a diachronic multi-source approach,
Proceedings of ICDM 2010 International Workshop on Mining Multiple Information Sources, Sydney, Australia, December 2010.

[Lamirel 2011] Lamirel, J.-C., Mall R., Cuxac P., Safi G.,
Variations to incremental growing neural gas algorithm based on label maximization,
Proceedings of IJCNN 2011, San Jose, CA, USA, August 2011.

[Falk ACL - 2012] Falk, I., Lamirel J.-C., Gardent C.,
Classifying French Verbs Using French and English Lexical Resources,
International Conference on Computational Linguistic (ACL 2012), Jeju Island, Korea, July 2012.

References

External references are in the proposed papers

[Lamirel 2012] Lamirel J.-C.,

A new diachronic methodology for automatizing the analysis of research topics dynamics : an example of application on optoelectronics research,

Scientometrics Special issue on 7th International Conference on Webometrics, Informetrics and Scientometrics and 12th COLLNET, Scientometrics 93(1): 151-166 (2012).

[Lamirel 2014] Lamirel J.-C., Falk I., Gardent C.,

Federating clustering and cluster labeling capabilities with a single approach based on feature maximization: French verb classes identification with IGNGF neural clustering,

Neurocomputing, Special issue on 9th Workshop on Self-Organizing Maps (WSOM 2012), 147, pp. 136-146, 2014.

[Lamirel – JADT 2014] Lamirel J.-C., Cuxac P.,

Une nouvelle méthode statistique pour la classification robuste des données textuelles : le cas Mitterand-Chirac,

JADT, Paris, France, April 2014.

References

External references are in the proposed papers

[Lamirel 2016a] Lamirel J.-C., Cuxac P., Hajlaoui K.,
A new approach for feature selection based on quality metric,
Advances in Knowledge Discovery and Management, 6 (665), Springer.

[Lamirel 2016b] Lamirel, J.-C., Dugué N., Cuxac P.,
New efficient clustering quality indexes,
Proceedings of IJCNN 2016, Vancouver, BC, Canada, July 2016.

[Dugué & Lamirel 2017] Dugue N., Lamirel J.-C., Tebbakh A.,
Feature selection and complex networks methods for an analysis of collaboration evolution in science: an application to the ISTEX digital library,
ISTE-ISKO Maghreb 2015 Post Proceedings Book.

[Al Zied 2018] Al Sied H., Dugue N., Lamirel J.-C,
Automatic summarization of scientific publications using a feature selection approach,
International Journal on Digital Libraries, 1-13 (2018).

References

External references are in the proposed papers

[Dugué et Lamirel 2018] Dugué N., Lamirel J.-C., Perez A.,

Bringing a feature selection metric from machine learning to complex networks,

Proceedings of Complex Networks 2018, 7th International Conference on Complex Networks and Their Applications, Cambridge, UK, September 2018.

[Olteanu et Lamirel 2019] Olteanu M., Lamirel J.-C.,

When clustering the multiscalar fingerprint of the city reveals its segregation patterns, Proceedings of 13th International Workshop on Self-Organizing Maps andLearning Vector Quantization, Clustering and Data Visualization,

Barcelona, Spain, June 2019.

[Lamirel 2020] Lamirel J.-C, Chen Y., Cuxac P., Al Shehabi S., Zeyuan L., Dugue N.,

Science of Science research in mainland China: 40 years of evolution. A new method of analysis based on clustering with feature maximization and contrast graphs,

Scientometrics (to appear).

Contact and questions

Please email me:

lamirel@loria.fr